

4) Develop a constrained generator of stereoisomers, (see Reference 9) including:

A) design and implement a complete and irredundant generator of possible conformations for a given known, or a candidate for an unknown, structure;

B) provide constraints for the conformation generator so that proposed structures for a known or unknown compound possess only those features allowed by: i) intrinsic structural features such as ring closure and dynamics of the chemical structure; and ii) data sensitive to molecular conformations (e.g., MCD, NMR);

C) integrate the stereochemical developments with the GENOA program as a final, comprehensive solution to the structure generation problem and allow for interface of the program with other methods dependent on atomic coordinates.

5) Promote applications of these new techniques to structural problems of a community of collaborators, including improved methods for structure elucidation and potential new biomedical applications, through resource sharing involving the following methods of access to our facilities and personnel;

A) nationwide computer network access, via the SUMEX-AIM computer resource;

B) exportable versions of programs to specific sites;

C) workshops at Stanford to provide collaborators with access to existing and new developments in computer-assisted structure elucidation in an environment where complex questions of utility and application can be answered directly by our own scientific staff;

D) interface to a commercially available graphics terminal for structural input and output, at as low a cost as possible, so that chemists can draw or visualize structures more simply and intuitively than with our current, teletype-oriented interfaces.

#### B. Justification and Requirements for Continued SUMEX Use

In previous sections we discussed the relationship between the DENDRAL Project and SUMEX-AIM, methods for using SUMEX-AIM for dissemination of our programs to a broad community of structural chemists and biochemists and a critique of resource management. In this section we wish to emphasize certain factors which were not discussed earlier and to show how our future directions and interests are closely related to the proposed continuation and augmentation of the SUMEX-AIM resource.

As resource-related research, DENDRAL is intimately tied to the SUMEX resource. Our involvement with SUMEX goes far beyond simple use of the facility. We use SUMEX as the focal point for a number of collaborative efforts, for export of our software and for the communication facilities essential to maintaining close contact with remote research groups working with us. SUMEX provides computational facilities for our workshops, where we bring outside investigators to Stanford to use new programs applied to real structural problems. We have already discussed in our critique the difficulties we have, in view of heavy SUMEX load, of maintaining both our research effort and the resource-sharing aspects of our project.

In view of these factors and because SUMEX is our sole source of computational facilities, we took certain steps in our renewal proposal to attempt to alleviate our situation. Specifically, we requested a computer for our own project, a DEC VAX 11/780, to be linked to SUMEX via ETHERNET. This computer was meant to help offload some of the computational burden DENDRAL places on SUMEX, to provide a facility for production use of our programs by our collaborators and to represent a model for the type of low-cost, scientific computer available in the future to many investigators who could then run our programs in their own laboratories.

Our request for the VAX was turned down with specific comments made that SUMEX facilities should be used to support development of new programs and to the extent possible, encourage preliminary production use of our programs by outside persons. In our opinion this view is somewhat shortsighted, because SUMEX is currently overloaded to the extent that even development is impeded. In addition, our current situation leaves no room for the computational burden created by some of our collaborators who need considerably more than "preliminary" access because they have no access to a computer suitable for running our programs.

For these reasons, we strongly support the effort of SUMEX to acquire a VAX and other small machines in future years. Although we realize that such machines will have to be shared among the SUMEX-AIM community as a whole, the augmentation of the resource would go a significant way to meeting the computational requirements of our project and provide a variety of systems of potential use for future export of our programs.

### C. Needs and Plans for Other Computing Resources

For several years now we have directed some attention toward alternative computing resources which could be used to support all "production" use of our programs, i.e., all applications designed to use the programs to solve real problems. Although this would have the severe disadvantage of separating our research effort from many of the applications, it has been our hope that emerging technology in networking would enable us to keep in reasonably close contact with another resource. Two resources have emerged as candidates for systems where our programs can be accessed and used in problem-solving. Unfortunately, neither has so far proven feasible for several reasons (mentioned below). At this time we cannot determine if the problems will be resolved. Until such time, we will remain completely dependent on SUMEX for all our computational needs.

One alternative resource is the NIH/EPA Chemical Information System. For more than three years we have been working with them to obtain sufficient contract money to provide a version of CONGEN integrated into that system. The concept and the funds were approved but a contract has never been issued due to administrative problems at the EPA. Although there have been some developments recently, we still have no firm idea on when such a contract will be issued. If this effort is successful, then we can encourage persons who desire access to our programs to consider using the NIH/EPA system.

A second alternative is the National Resource for Computation in Chemistry (NRCC). This Resource has recently had its funding terminated.

We are now pursuing an alternative discussed previously, that of arranging license agreements with private industry for dissemination of our software. This will likely be the focus of our future efforts to disseminate programs to those researchers who merely wish to use them rather than work together with us in collaborative arrangements to develop more powerful programs.

#### D. Recommendations for Future Resource and Community Development

We have discussed previously our recommendation for the hardware augmentation, particularly with regards to purchase of small machines to facilitate future export. We also have increasing need for more file storage on-line. This is a result of building large data bases as part of our research in spectral interpretation. For the time being we are working with experimental programs and small data bases. As time progresses, however, these data bases will grow rapidly as our group and a number of our collaborators add additional structures and associated spectral data.

Another capability which is of increasing importance to our own work is access to low-cost graphics systems. Our programs will develop increasing dependence on graphics for visualization of three-dimensional molecular structures. Scientists desiring access to our programs will need a graphics terminal for optimum use of our systems. Currently available vector displays are simply too expensive for the average investigator. The emerging technology of low-cost raster display systems offers a more promising possibility. However, no currently available machine has the required capabilities for under \$10,000, and this is an area where machines like the Alto hold more promise. SUMEX could perhaps initiate an effort to obtain a system which has the hardware necessary for frame-based display. Such a system allows rotation of three-dimensional objects in a way which permits visualization of the actual shape of the object.

II.A.1.4 EXPEX Project

## EXPEX - Expert Explanation Project

Edward H. Shortliffe, M.D., Ph.D.  
Departments of Medicine and Computer Science  
Stanford University

Michael R. Genesereth, Ph.D.  
Computer Science Department  
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

## A. Project Rationale

EXPEX is a new Stanford project that joined the AIM community only a few months before this report was prepared. We therefore have little to report in the way of progress other than our background work that led to a recently funded proposal and the initiation of this new research. The major thrust of the work is the development of powerful representational schemes to facilitate knowledge acquisition and explanation. This includes not only the study of fundamental representational formalisms but also the encoding of various types of knowledge, such as causal information and user models.

We believe that the productivity of basic computer science research tends to be heightened by experiments that deal with significant real world problem domains. Challenges drawn from chemistry, medicine, and molecular biology have introduced additional complexity to expert systems work at Stanford, but have simultaneously forced system developers to respond to pragmatic constraints and user demands that have had a significant impact on the basic AI techniques selected or developed. Thus, we believe that creative investigation into symbolic reasoning techniques is facilitated by working in real world settings where the application forces us to avoid oversimplification. The explanation portion of the research effort will therefore deal with a medical domain (endocrinology) and be undertaken on SUMEX, whereas the knowledge acquisition portion will deal with nonmedical topics and use other computing resources at Stanford. Our report here will only describe EXPEX, the research on expert medical explanation, but it should be understood that this is actually only one part of a coordinated effort to tie together research in knowledge acquisition and explanation through common representation techniques.

## B. Medical Relevance and Collaboration

Our interest in explanation derives from the insights we gained in developing explanatory capabilities for the MYCIN system. In the case of MYCIN and its descendents, we have been able to generate intelligible explanations by taking advantage of its rule-based representation scheme. Rules can be translated into English for display to a user, and their

interactions can also be explicitly demonstrated. By adding mechanisms for understanding questions expressed in simple English, we were able to create an interactive system that allowed physicians to convince themselves that they agreed with the basis for the program's recommendations. The limitations of the explanations generated in this way have become increasingly obvious, however, and have led to improved characterization of the kinds of explanation capabilities that must be developed if clinical consultation systems are to be accepted by physicians.

### C. Highlights of Research Progress

MYCIN's explanation capabilities were generalized in EMYCIN and thus became available for any EMYCIN consultation system. They were further modified and utilized in both TEIRESIAS and GUIDON. Although we had experienced problems using MYCIN's rules for certain kinds of explanations (e.g., control mechanisms that were sometimes encoded in rules, or algorithmic knowledge such as the mechanisms for drug selection), it was in the setting of GUIDON that the inadequacies of MYCIN's approach became most apparent. Consider, for example, a simple MYCIN rule such as:

```
If:      the patient is less than 8 years old
Then:    don't give tetracycline
```

This rule is adequate for MYCIN's decision making task, and would be understood by most physicians if it were used in an explanation, but it is obvious to a casual observer that it contains a giant leap in logic. It is accordingly difficult for GUIDON to teach this rule to a novice medical student because the underlying pathophysiologic knowledge (i.e., that tetracycline is deposited in the developing bone and teeth of youngsters, weakening the former and disfiguring the latter) is not explicitly represented in MYCIN. Examples such as this one emphasize that a variety of knowledge forms are necessary if an intelligent system is to customize its explanations to the individual who is using the program. Underlying structural and causal relationships are generally required in addition to the high level judgmental rules that had contained almost all of the domain knowledge in MYCIN and the other EMYCIN systems.

We therefore began to study in more detail the nature of the explanatory process, and were surprised to find that there are very few writers who have addressed the issues which now interest us. Perhaps the most relevant studies of explanation are in the education literature; several educators have tried to identify the characteristics of explanations which make individuals good teachers. These analyses are accordingly relevant to computer-aided instruction work, such as GUIDON, although issues of automation are not addressed explicitly. On the other hand, they seem less pertinent when applied to the "persuasiveness" of a justification offered by a scientist to a colleague.

A weekly seminar group has been formed to discuss knowledge representation and to analyze the characteristics of good explanations. We have often kept our discussions separate from computer science issues, concentrating instead on the psychology of explanation and planning to

return eventually to consider ways in which our developing theory might be implemented in knowledge-based consultation systems. Although there are several subproblems, it was agreed that the problems of explanation can generally be divided into four categories: (1) modeling the knowledge of the system user; (2) selecting a response strategy; (3) modeling contextual information regarding the interaction; and (4) understanding the question. One goal of our new work, then, is to build an explanation system which explicitly addresses these topics.

#### Modeling the User's Knowledge

GUIDON and other ICAI systems have recognized the need to keep an internal model of the student, i.e., what he has shown he knows, what you have already told him, and perhaps a record of where his greatest weaknesses lie. Similarly, it is clear that an expert human consultant customizes his explanations so that they can be understood by the person requesting the consultation (and are thereby maximally convincing). The expert starts with certain suppositions about his client's knowledge (e.g., a teacher may presume his student is starting from scratch, but a cardiologist will assume that another physician requesting advice probably already knows a fair amount of cardiology). The default presumption is modulated, however, as the interaction proceeds and the client demonstrates his strengths or weaknesses.

We have recently begun some experiments to investigate methods for encoding, along with the domain knowledge, the complexity and importance of that knowledge. These two parameters seem to be independently important in deciding whether to include a given reasoning step in an explanation. "Key" points (i.e., those that are highly important) probably should be mentioned even if they are not complex and are likely to be known to the user. On the other hand, less important but complex items probably need not be mentioned unless an expert user is really pressing for details of a decision pathway. Thus, static measures of complexity and importance can be compared with user descriptors that are initially assigned by default (depending upon the status of the user, e.g., expert vs. student), but are later altered dynamically in response to the course of the dialog and what it has revealed about the user's background knowledge.

These ideas have been encoded in a small computer program which uses a limited knowledge base of rules and associations from the domains of pharyngitis (sore throats) and calcium metabolism. We have experimented with a semantic network representation in which the nodes are values of attributes and rules are only one form of link between nodes. All nodes and rules have complexity and importance measures associated with them. An "opinion" regarding a specific patient can be represented as a subset of the nodes in the network, plus the links between them that account for how it has been determined which nodes are active. In this setting, a question tends to ask how it has been determined that a given node is active for a given patient. The appropriate explanation could be very complex if an effort were made to explain every link leading from data observations to the node descriptor in question. A customized explanation is therefore generated based on three variables which can be dynamically manipulated by the program: (1) the focus of the dialog (e.g., broad-based vs. localized),

(2) the expertise of the user, and (3) the degree of generality which is appropriate. These three variables are clearly not independent, and we are experimenting with ways to have their values manipulated in a reasonable fashion as the dialog proceeds.

This early effort has provided the basis for further discussions in our seminar group as we have attempted to arrive at an optimal representation for the research to follow. We have been fortunate to enlist the collaboration of an endocrinologist at Stanford, Dr. Larry Crapo, who is eager to work with us on building an endocrinology knowledge base. It is likely that we will select the pathophysiology of calcium disorders as a small focused area to study. This domain is appealing for computer-based representation because the relationships are well-understood and there are some challenging problems of feedback homeostasis that will need to be represented. In the years ahead, we will encode this knowledge base in detail and begin experiments on the generation of explanations using the kinds of techniques outlined above.

#### Selecting A Response Strategy

Our explanation efforts to date have tended to be simple reiterations of individual reasoning steps, but it is clear that experts and teachers use several alternate strategies for conveying their ideas or key facts. Many of these techniques draw upon common sense world knowledge (e.g., analogies with familiar concepts outside the domain), but we have thus far failed to capitalize on these teaching strategies in our work. Thus another goal of the work that lies ahead will be to develop structures for drawing parallels or otherwise representing the strategies used by good "explainers."

#### Modeling Contextual Information Regarding the Interaction

We have already mentioned some of the ways in which contextual information may be useful in determining the best way to answer a question. For example, a more accurate model of the user's knowledge can be developed over time, and the extent to which a given conversation is focused on a particular local topic can be assessed. Note that we are emphasizing here issues other than those related to natural language understanding; computational linguists also often cite the need to record contextual dialog information in order to handle problems such as anaphora. An understanding of the "flow" of a dialog is also important in understanding the meaning of subsequent questions, as we discuss below.

#### Understanding The Question

This issue interfaces with the problem of natural language understanding, but we view it in a somewhat different light. We emphasize instead the ways in which the model of the user and contextual information may allow us to disambiguate questions. To draw from a medical example that we have frequently discussed, consider the following scenario. A reasoning program for pharyngitis diagnosis and management has just diagnosed strep throat and recommended penicillin and the user asks the question "Why would you give penicillin?" In the most obvious case, one

might imagine a response that itemizes the risks of streptococcal infections and the reasons for treating early with penicillin. Similarly, one might expect a more detailed response for a student and a quick summary for a physician using the system.

However, an alternate interpretation is that EVERY physician knows the theoretical reasons for giving penicillin in strep pharyngitis, and that if the user is a physician and is asking the question then he must be asking something different than the simple informational question. In this case the query might be interpreted as a challenge (one that might have been conveyed by tone of voice if it had been asked of a human consultant). Apparently the user has reason to doubt that penicillin was the appropriate agent in this case, or thinks that no drug was required. Other background information and contextual knowledge should also help, and an intelligent program might thereby answer the question in a given case in any of the following ways:

"Because the patient has pre-existing rheumatic heart disease."

"Because I doubt that he is allergic to penicillin, even though he reported that he is."

"Because he is unreliable and I am afraid I will not be able to reach him to call him back if his strep culture comes back positive."

"Because I tend to treat conservatively and give penicillin for strep throat even though I know there hasn't been a case of rheumatic heart disease in California in over 10 years."

Note how different these kinds of explanations are from the simple justification that a program such as MYCIN might have given:

"Because streptococcal pharyngitis may be followed by rheumatic myocarditis or glomerulonephritis, mediated by immune complexes, and I can prevent this complication by giving penicillin (to which streptococci are uniformly sensitive)."

The ideal intelligent assistant should be able to determine from knowledge of the user, the domain, the individual case, and the context of the dialog, which of the preceding responses is most appropriate. We will attempt to identify methods for giving our program this kind of capability.

#### D. Publications Since January 1980

Wallis, J.W. and Shortliffe, E.H. Explanatory power for expert systems: studies in the representation of causal relationships for medical consultations. Internal working memo, Heuristic Programming Project, Stanford University, May 1981.



### E. Funding Support

Grant Title: "The Development of Representation Methods to Facilitate Knowledge Acquisition and Exposition in Expert Systems"  
Principal Investigator: Edward H. Shortliffe  
Agency: Office of Naval Research  
ID Number: NR 049-479  
Term: January 1981 to December 1983  
Total award: \$456,622  
Current award (1981): \$140,825

## II. INTERACTION WITH THE SUMEX-AIM RESOURCE

### A. Medical Collaborations and Program Dissemination via SUMEX

We are only beginning program development at this time, and have therefore had no opportunity to share our results with others as of yet.

### B. Sharing and Interaction with Other SUMEX-AIM Projects

We anticipate frequent ongoing interactions with other SUMEX-AIM research efforts because the development of explanation techniques is a pertinent research issue for all expert systems work in medicine. Bill Clancey's work on GUIDON is addressing many of the same issues and we expect frequent opportunities for interchange.

### C. Critique of Resource Management

Although we have not yet placed significant demands on SUMEX management, our previous experience working with Tom Rindfleisch and his staff would suggest that this new project will receive the same kind of laudatory service for which SUMEX has become known.

## III. RESEARCH PLANS (6/81-12/83)

### A. Project Goals and Plans

We intend to investigate optimal techniques for the computer-based representation of expert knowledge. Because we have come to recognize the limitations of any single representation technique taken alone, a principal objective will be to merge alternate approaches, augmented with new capabilities. We will, in turn, evaluate the effectiveness of the new representation scheme by focusing on issues of both knowledge acquisition and explanation. Furthermore, we will perform these experiments in two expert domains, medical reasoning (EXPEX) and computer circuitry debugging (DART). These areas were selected because we have local expertise in each, but also because they are sufficiently different from one another that they will force us to ensure the generality of the techniques we are developing. Utilizing a single representation scheme for all aspects of the work will also encourage generality of the developed techniques because this will

force us to avoid concentrating on either the input (knowledge acquisition) or output (explanation) functions alone.

Initially we shall concentrate on defining the knowledge representation scheme to be used. Although modifications will of course be necessary in response to additional lessons learned thereafter, we expect to reach an early consensus on the major components of the internal representation we will be using.

Subsequently our efforts will divide into two components, each of which will utilize the representation scheme devised in the initial period. EXPEX will concentrate on manually constructing a knowledge base regarding calcium metabolism and pathophysiology, whereas the DART effort will be concentrating on knowledge acquisition for their non-medical domain. In the EXPEX work, the construction of the clinical knowledge base will have created an environment for the development of the explanation capabilities which are the second thrust of our work. Drawing on the early work of Jerry Wallis, described in the memo referenced in the publications section of this report, we will next construct an expository system that uses the endocrinology knowledge base in order to generate interactive explanations.

Ultimately we hope to perform experiments using both the knowledge acquisition and explanation tools that will have been developed in the two separate domains. One task will be to see if we can develop from scratch the endocrinology knowledge base that will have been hand-coded for the EXPEX effort. Because this knowledge will have previously been encoded manually, we will have a well-defined model of the form the knowledge should take as it is acquired interactively using the new system building tools developed in the electronics environment. At the same time, some of us will be developing experiments to test the validity and effectiveness of the explanation tools that were developed for EXPEX. An excellent test of the generality we have been seeking will be to build the circuit debugging knowledge base using the new knowledge acquisition tools, and then to demonstrate the utility of the explanation routines for exposition of the knowledge in this new domain.

Our ultimate goal, then, is to have developed a unified system of knowledge representation that facilitates both system building, through interactive knowledge acquisition, and explanation, through interactive responses to knowledge base queries. It should be emphasized that throughout our work the focus will be on the underlying representation issues and not on polished text generation nor natural language understanding.

II.A.1.5 MOLGEN Project

## MOLGEN - A Computer Science Application to Molecular Biology

Prof. E. Feigenbaum and Dr. P. Friedland  
Department of Computer Science  
Stanford University

Assoc. Prof. D. Brutlag  
Department of Biochemistry  
Stanford University

Assoc. Prof. L. Kedes  
Department of Medicine  
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

## A. Project Rationale

The MOLGEN project has focused on research into the applications of symbolic computation and inference to the field of molecular biology. This has taken the specific form of systems which provide assistance to the experimental scientist in various tasks, the most important of which have been the design of complex experiment plans and the analysis of nucleic acid sequences. We plan to expand and improve these systems and build new ones to meet the rapidly growing needs of the domain of recombinant DNA technology. We do this with the view of including the widest possible national user community through the facilities available on the SUMEX-AIM computer resource.

It is only within the last few years that the domain of molecular biology has needed automated methods for experimental assistance. The advent of rapid DNA cloning and sequencing methods has had an explosive effect on the amount of data that can be most readily represented and analyzed by computer. Moreover we have already reached a point where progress in the analysis of the information in DNA sequences is being limited by the combinatorics of the various types of analytical comparison methods available. The application of judicious rules for the detection of profitable directions of analysis and for pruning those which obviously lack merit will have an autocatalytic effect on this field in the immediate future.

The MOLGEN project has continuing computer science goals of exploring issues of knowledge representation, problem-solving, and planning within a real and complex domain. The project operates in a framework of collaboration between the Heuristic Programming Project (HPP) in the Computer Science Department and various domain experts in the departments of Biochemistry, Medicine, and Genetics. It draws from the experience of several other projects in the HPP which deal with applications of artificial intelligence to medicine, organic chemistry, and engineering.

During the next three years of MOLGEN research we intend to begin a transition from being primarily a computer science research project to being an interdisciplinary project with a strong applications focus. The tools that we have already developed will be improved to the point where they make a significant contribution to both research and engineering in the domain of molecular biology.

## B. Medical Relevance and Collaboration

The field of molecular biology is nearing the point where the results of current research will have immediate and important application to the pharmaceutical and chemical industries. Recombinant DNA technology has already demonstrated the possibility of harnessing bacteria to produce nearly limitless amounts of such drugs as insulin and somatostatin. Governmental reports estimate that there are more than 200 new and established industrial firms already undertaking product development using these new genetic tools.

The programs being developed in the MOLGEN project have already proven useful and important to a considerable number of molecular biologists. Currently several dozen researchers in various laboratories at Stanford (Prof. Paul Berg's, Prof. Stanley Cohen's, Prof. Laurence Kedes', Prof. Douglas Brutlag's, Prof. Henry Kaplan's, and Prof. Douglas Wallace's) and over 300 others throughout the country are using MOLGEN programs over the SUMEX-AIM facility. We have exported some of our programs to users outside the range of our computer network (University of Geneva [Switzerland], Imperial Cancer Research Fund [England], and European Molecular Biology Institute [Heidelberg] are examples).

## C. Highlights of Research Progress

### Accomplishments:

The current year has seen the completion of what might be considered the first phase of the MOLGEN project. This section will summarize the major accomplishments of that first phase.

### 1. Representation Research

The domain of molecular biology has proven a fruitful testbed in the development of a flexible software package, the Unit System, for symbolic representation of knowledge. The package is already in use by a variety of research projects both within the Heuristic Programming Project at Stanford and at other institutions. It provides for acquisition and storage of many different types of knowledge, ranging from simple declarative types like integers and strings to complex declarative types like nucleic acid restriction maps to procedural types like a rule language in a subset of English.

A major effort has been made in the past year to take the Unit System and, observing the experience of its many scientific users, improve and enhance those features which are most important. This has resulted in a speed improvement of at last two orders of magnitude for the most used

functions in the representation system. The MOLGEN project has provided a unique laboratory for the conversion of a theoretically-based knowledge base system into a practical package for knowledge acquisition and manipulation. This is because of the active daily use of the system for real laboratory problems. The Unit System has become what may be considered the first "second-generation" knowledge representation package.

We have concentrated on representation methods that are unique to molecular biology, particularly, convenient methods for storing information about nucleic acid sequences and maps of those sequences, as well as an English-like language for manipulating that information. This language has allowed and encouraged the molecular biologist members of the MOLGEN project to become their own "programmers" without having to worry about the underlying representation structures of their knowledge bases. For example, the phrase:

```
JOIN SV40 FROM FIRST ECOR1 SITE TO FIRST BAMH1 SITE TO PBR322 FROM 250  
TO 3000 INTO NEWSEQUENCE
```

will perform the clearly indicated operation.

## 2. Planning Research

The problem of designing laboratory experiments in molecular biology has been fundamental to MOLGEN research. The work has been split into two major subparts, each resulting in a doctoral thesis in computer science. The two systems, developed by Peter Friedland and Mark Stefik, produce reasonable experiment designs on test problems suggested by laboratory scientists.

The majority of MOLGEN planning work has awaited the successful completion of the latest phase of representation improvement described above. Dr. Rene Bach, a post-doctoral fellow, has recently completed a DNA sequencing experiment adviser, built entirely with the Unit System and associated procedural description language. This system provides guidance in developing a sequencing protocol given a partial initial restriction map for a new nucleic acid sequence. Other members of the MOLGEN group have begun exploring the representations needed with the Unit System to provide transparent descriptions (i.e., invisible to the non-computer scientist user) of growing experimental plans, so that plans may be treated in a manner identical with all other types of knowledge.

## 3. Knowledge Base Construction

Over six man-years have now been spent in constructing knowledge bases for various fields of molecular biology. Professors Kedes and Brutlag, Dr. Bach, and several students have cooperated on one knowledge base which is expert in restriction enzyme methodology and another which is competent for a wide range of general laboratory techniques. Professor Kedes has worked on a knowledge base for his own interests in gene structure. Professor Brutlag is concentrating on a knowledge base for satellite DNAs. Dr. Bach has built a knowledge base expert in sequencing methods for his sequencing advisor (see above). Professor Sninsky and Dr.

Bach have collaborated in a knowledge base for expression vectors. Professor Sninsky and Dr. Abarbanel have collaborated on the beginnings of a protein knowledge base to explore methods for predicting secondary protein structure from primary amino acid sequence. Several researchers in Professor Hogness's laboratory in the Department of Biochemistry have begun to build a knowledge base for storing information about many different lambda vector clones. Finally, several scientists in Professor Kaplan's Cancer Biology Research Laboratory have started to explore a monoclonal antibody knowledge base.

The knowledge bases developed under the MOLGEN grant have begun to find their way into the daily laboratory practice of many of the scientists associated with the project. They have provided a mechanism for managing the explosive growth of data and strategies in many areas of molecular biology without the necessity of building special purpose systems for each area. Also, the expert scientists themselves have been able to design and build their own systems, avoiding the time and reliability problem of a knowledge base passing through the filter of a computer scientist intermediary. The knowledge bases have served as "intelligent encyclopedias," as simulation systems, and as training vehicles.

It should also be noted that the Unit System allows for the easy transfer of knowledge from one knowledge base to another, and indeed the various expert molecular biologists have freely shared information as they work on related knowledge bases.

#### 4. Other Applications of Symbolic Computation to Molecular Biology

MOLGEN programmers have spent full time enhancing existing MOLGEN applications programs and developing new systems. The SEQ program is a general purpose nucleic acid sequence analysis system. It provides a range of functions including translation, lexicography, regions of richness, restriction mapping, general string search, and intra- and inter-sequence homologies, symmetries, and dyad symmetries. The program calculates statistical probabilities for homologies, symmetries, and dyad symmetries, and also determines approximate free energy contribution of dyad symmetry structures in RNA. SEQ is highly interactive and provides many built-in explanation and help facilities.

MAP is a program which determines restriction sites from enzymatic digest data. A recent collaboration with Dr. Pearson has been exploring ways to combine the ideas from MAP and Dr. Pearson's system for solving similar problems.

Work has begun on two major new applications programs: GEL, a system which provides bookkeeping and overlap determining assistance for "shotgun" sequencing experiments; and AA a program which provides most of the functions of SEQ for amino acid sequences.

During the first year of the current MOLGEN grant, we have provided guest access to the SEQ and MAP programs to the national academic community through the facilities of the SUMEX-AIM computer system. This has meant free, dial-up access from almost anywhere in the United States. Over 300

researchers at over 80 institutions have used the service. It has been so popular that the SUMEX-AIM Executive Committee has found it necessary to limit the service to at most two simultaneous users at any one time. The facilities provided to MOLGEN guests has been very limited, with a single directory and 250 disk pages serving the entire national community. Despite this, a wide variety of interesting research has been done, and MOLGEN is most grateful to the SUMEX-AIM staff for their kind assistance.

#### Research in Progress:

The remainder of the current grant period will be spent on the further development of the tools that have been constructed for experiment design and sequence analysis and on expansion and improvement of the knowledge base. This section details those research plans.

#### 1. Representation

The Unit System is now at a stage of general utility. The MOLGEN group will continue to enhance and improve both the underlying representation methods and the user interface. A further order of magnitude improvement in the speed of common operations will be achieved during the next year. There will be at least a doubling of the genetic-specific vocabulary of the procedural description language. Mechanisms will be developed to allow the Unit System to communicate effectively with all of the major MOLGEN applications programs, particularly SEQ, MAP, and GEL.

We also anticipate a major effort, beginning in the summer of 1981, to adapt the Unit System to run on one of the newly available personal scientific work-station computers--most likely the Xerox Dolphin. This will provide a qualitative improvement in user interaction because of the large bit-map display and graphics capabilities, and the ability to free individual knowledge base builders from time-sharing system load.

#### 2. Knowledge Base Development and Planning

Planning work will proceed by moving into at least one new sub-domain and by synthesizing the research of previous years. Dr. Bach will construct a knowledge-based system for the planning of cloning experiments using the experience he has gained from the simpler task of designing sequencing experiments. The previous MOLGEN planning research has produced two major ideas. One, the "skeletal-plan" approach of Dr. Friedland, involves the selection and refinement of planning strategies, which may range from abstract to specific, provided as part of an expert knowledge base. This idea resulted from a study of the way in which molecular biologists design experiments. The other idea was the "constraint-posting" method of Dr. Mark Stefik, which concentrates on the extensive evaluation of the constraints introduced by each step of a growing plan in order to guide the selection of the next step of the plan. The MOLGEN group will undertake the project of synthesizing these two ideas, combining the practical efficiency of skeletal-plan refinement with the general power of constraint posting.

We will also begin work on plan verification and optimization systems within the same general framework; i.e. systems to check proposed plans for suitability and to improve such plans. We hope to begin to extend this to a plan debugging system; one which interacts with an experimenter to determine where an experiment failed and how to correct the problem.

Knowledge base construction will proceed, with each MOLGEN collaborator exploring the particular sub-domain most interesting to him. Professor Maxam will begin building a knowledge base, most likely in the area of detailed nucleic acid structure, this summer. This will lead to joint work with Dr. Friedland on the application of knowledge-based methods to the study of the mechanism of gene regulation. Several researchers at the Imperial Cancer Research Fund in London, England will join the effort starting in the early autumn of 1981.

In particular, we expect the area of cloning methodology to occupy a large portion of Dr. Bach's time in knowledge base construction. We will draw from the wide and varied expertise within the Stanford community during this effort.

### 3. Applications Systems

Work will continue on improving and enhancing SEQ and MAP. The GEL and AA programs will begin operation. The currently available tools within the Unit System allow the molecular biologists themselves to design and construct special-purpose applications systems within their knowledge bases. Professors Brutlag and Kedes have already built over a dozen such "programs" for construction of specific restriction maps and fragment tables and the simulation of recombinant DNA operations. As the domain experts decide that these systems are generally useful, they will be optimized and packaged into stand-alone programs.

The MOLGEN group will continue to cooperate with any national efforts to develop a sequence analysis data bank and facility for the academic community. We hope our current collaborative service activities on SUMEX-AIM will serve as the prototype for a larger and more comprehensive national facility.

### D. Publications

Feitelson J., Stefik M.J., A Case Study of the Reasoning in a Genetics Experiment, Heuristic Programming Project Report HPP-77-18 (Working Paper) (May 1977)

Friedland P., Knowledge-Based Experiment Design in Molecular Genetics, Proceedings Sixth International Joint Conference on Artificial Intelligence, 285-287 (August 1979)

Friedland P., Knowledge-Based Experiment Design in Molecular Genetics, Ph.D. Thesis, Stanford CS Report CS79-760 (December 1979)

Martin N., Friedland P., King J., Stefik M.J., Knowledge Base Management for Experiment Planning in Molecular Genetics, Fifth International Joint Conference on Artificial Intelligence. 882-887 (August 1977)



Stefik M., Friedland P., Machine Inference for Molecular Genetics: Methods and Applications, Proceedings of the National Computer Conference, (June 1978)

Stefik M.J., Martin N., A Review of Knowledge Based Problem Solving As a Basis for a Genetics Experiment Designing System, Stanford Computer Science Department Report STAN-CS-77-596. (March 1977)

Stefik M., Inferring DNA Structures From Segmentation Data: A Case Study, Artificial Intelligence 11, 85-114 (December 1977)

Stefik, M., An Examination of a Frame-Structured Representation System, Proceedings Sixth International Joint Conference on Artificial Intelligence, 844-852 (August 1979)

Stefik, M., Planning with Constraints, Ph.D. Thesis, Stanford CS Report CS80-784 (March 1980)

#### E. Funding Support

The MOLGEN grant is titled: MOLGEN: A Computer Science Application to Molecular Biology. It is NSF Grant ECS-8016247. Current Principal Investigators are Edward A. Feigenbaum and Bruce G. Buchanan, Professors of Computer Science, Laurence H. Kedes, Investigator, Howard Hughes Medical Institute and Associate Professor of Medicine, and Douglas L. Brutlag, Associate Professor of Biochemistry. MOLGEN is currently funded from 10/80 to 9/81 at \$146,582 including indirect costs as the first year of a three year renewal.

## II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

Until this year, all system development had taken place on the SUMEX-AIM facility. The facility has not only provided excellent support for our programming efforts but has served as a major communication link among members of the project. Systems available on SUMEX-AIM such as INTERLISP, TV-EDIT, and BULLETIN BOARD have made possible the project's programming, documentation and communication efforts. The interactive environment of the facility is especially important in this type of project development.

Unfortunately, the computing environment at SUMEX has suffered in the recent past from heavy demands on cycle time creating serious real-time delays for programmers and knowledge-base building especially. The Units Editor is especially sensitive because of its relatively large demands on cpu-resources. Accordingly, a significant fraction of the MOLGEN group activity has been transferred to the SCORE computer in the Department of Computer Science at Stanford. When SUMEX hardware is updated, we anticipate that its response time will improve and the MOLGEN computing will return full time to SUMEX. It is clear, however, that the MOLGEN project continues to thrive and prosper because of the computing environment only available at SUMEX: the interactive environment including instantaneous communications among collaborators who are physically distant (even on the Stanford campus), and especially the unique telecommunications

facilities that have allowed the development of the GENET community with its access to MOLGEN applications tools are two clear examples.

We have taken advantage of the collective expertise on medically-oriented knowledge-based systems of the other SUMEX-AIM projects. In addition to especially close ties with other projects at Stanford, we have greatly benefitted by interaction with other projects at yearly meetings and through exchange of working papers and ideas over the system.

The ability for instant communication with a large number of experts in this field has been a determining factor in the success of the MOLGEN project. It has made possible the near instantaneous dissemination of MOLGEN systems to a host of experimental users in laboratories across the country. The wide-ranging input from these users has greatly improved the general utility of our project.

We find it very difficult to find fault with any aspect of the SUMEX resource management. It has made it easy for us to expand our user group, to give demonstrations (through the 20/20 adjunct system), and to disseminate software to non-SUMEX users overseas.

### III. RESEARCH PLANS

#### A. Justification and Requirements for Continued SUMEX Use

The MOLGEN project depends heavily on the SUMEX facility. We have already developed several useful tools on the facility and are continuing research toward applying the methods of artificial intelligence to the field of molecular biology. The community of potential users is growing nearly exponentially as researchers from most of the bio-medical fields become interested in the technology of recombinant DNA. We believe the MOLGEN work is already important to this growing community and will continue to be important. The evidence for this is an already large list of pilot exo-MOLGEN users on SUMEX.

SUMEX is currently having difficulty meeting the research needs of the MOLGEN project adequately. We expect to need more file space as our knowledge bases grow; perhaps an additional 5000 disk blocks in the next few years for that work. Our real difficulties will come in the applications testing of MOLGEN tools. We support with great enthusiasm the acquisition of satellite computers for technology transfer and hope that the SUMEX staff continue to develop and support these systems. One of the oft-mentioned problems of artificial intelligence research is exactly the problem of taking prototypical systems and applying them to real problems. SUMEX gives the MOLGEN project a chance to conquer that problem and potentially supply scientific computing resources to a national audience of bio-medical research scientists.

II.A.1.6 MYCIN Projects Group

## MYCIN Projects

Edward H. Shortliffe, M.D., Ph.D.  
Departments of Medicine and Computer Science  
Stanford University

Bruce G. Buchanan, Ph.D.  
Computer Science Department  
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

## A. Project Rationale

The MYCIN Projects are a set of related research programs, each devoted to the development of knowledge-based expert systems for application to medicine and the allied sciences. The name was derived from our first system, the MYCIN program. That research has now given way to three active sub-projects (EMYCIN, GUIDON, and ONCOCIN), each of which is discussed in the sections that appear below. The key issue for all sub-projects has been to develop programs that can provide advice similar in quality to that given by human experts, and to develop systems that are easy to use and acceptable to physicians and medical students.

The success of the original MYCIN infectious disease consultation program has led us to try to generalize and expand the methods employed in that program to a number of ends:

- (1) to develop consultation systems for other domains (our generalized system-building tool is known as "Essential MYCIN", or EMYCIN, and has been applied in several new areas; ONCOCIN is our newest consultation system and was inspired by EMYCIN although it is actually an entirely new program);
- (2) to explore other uses of the MYCIN knowledge base (our tutoring system, GUIDON, uses the infectious disease knowledge in MYCIN to teach medical students about diagnosis and management of infections);
- (3) to continue to improve the interactive process, both for the developer of a knowledge-based system, and for the user of such a system (both EMYCIN and ONCOCIN have stressed simplified techniques for interacting with a knowledge base and entering data); and
- (4) to experiment with alternate techniques for knowledge representation, recognizing that the pure production rule method used in MYCIN was inadequate at times and frequently led to confusion regarding the separation of strategy or control

processes from domain knowledge (ONCOCIN uses production rules as only one of several knowledge representation techniques, and the work on GUIDON has led to a more robust revised version of MYCIN known as NEOMYCIN).

## B. Medical Relevance and Collaboration

By utilizing our EMYCIN system to collaborate on building the PUFF program, we learned that it is possible in a short period of time to develop a clinically useful consultation system using the domain-independent parts of MYCIN. EMYCIN has since been applied in a number of additional medical domains. With each successive application we learn more about the representation of medical knowledge and the scope and limitations of the production rule formalism used in EMYCIN. For example, it has become clear that "shallow" rules relating signs and symptoms to diagnoses through a few intermediate concepts can be sufficient for high performance in medical diagnosis. On the other hand, such shallow rules are not always sufficient for teaching medical students because they lack the deeper causal links needed for justifying and remembering the more shallow associations (see GUIDON discussion below).

Although EMYCIN was not used to build our new ONCOCIN program, the lessons learned in building prior production rule systems have allowed us to create a large oncology protocol management system much more rapidly than was the case when we started to build MYCIN. We are introducing ONCOCIN for use by Stanford oncologists in the Spring of 1981. This would not have been possible, of course, without the active collaboration of Stanford oncologists who helped with the construction of the knowledge base and also kept project computer scientists aware of the psychological and logistical issues related to the operation of a busy outpatient clinic.

In addition, there is a growing realization that medical knowledge, originally codified for the purpose of computer-based consultations, may be utilized in additional ways that are medically relevant. Using the knowledge to teach medical students is perhaps foremost among these, and GUIDON continues to focus on methods for augmenting clinical knowledge in order to facilitate its use in a tutorial setting. A particularly exciting aspect of this work is the insight that has been gained regarding the need to structure knowledge differently, and in more detail, when it is being used for different purposes (e.g., teaching as opposed to clinical decision making). This aspect of the GUIDON research has led to the development of a modified version of MYCIN, NEOMYCIN, which is an evolving computational model of medical diagnostic reasoning that we hope will enable us to better understand and teach diagnosis to students.

## C. Highlights of Research Progress

### 1. Accomplishments This Past Year

#### EMYCIN

In the last year, substantial research efforts were completed and described in publications. First, a complete EMYCIN system, including a

rule compiler and interactive rule editor, was packaged and documented in Bill van Melle's thesis. Second, an investigation into mixing production rules and frames as a representation of medical knowledge was completed and documented by Jan Aikins. Third, a redesign of the EMYCIN system to include more of the structural and strategic knowledge needed for tutoring was completed and documented (see description of NEOMYCIN in the GUIDON sections below). In addition, the LISP code and the EMYCIN Manual were both improved considerably in response to suggestions from outside users.

The complete EMYCIN system includes: a) a rule interpreter, b) an explanation facility, c) an abbreviated rule language and editor for rule input, d) a debugging package, and e) a rule compiler. Some new developments in these parts are described briefly below and details are given in van Melle's thesis.

The rule interpreter has remained much the same since its conceptualization as a procedure that traces backward through chains of rules asking questions of the user only when values of parameters cannot be deduced. One important new development was a change in the certainty factor model used to propagate degrees of certainty from multiple pieces of positive and negative evidence to a conclusion. The new model is commutative, which means that it is no longer necessary to accumulate separate measures for positive and negative evidence, but a single measure reflecting both. The new model gives the same result as the old one in combining certainty factors of like sign, but is more gentle when combining CF's of opposite sign. The previous scheme had the problem of compressing too much information into the region near 1, and, as a result, ten conclusions of  $CF = 0.9$  could be substantially overthrown by a single conclusion of  $CF = -0.8$ .

The explanation capability was improved by implementing the system's dictionary in hash tables. This gives the reasoning program more working space, since the dictionary can be kept on secondary storage until needed. Also, access time for parts of the dictionary is small because of the hash coding. The mechanisms used in EMYCIN have been used in other systems as well.

The input language for new rules has been simplified and stylized in an abbreviated rule language (ARL). ARL exploits the fact that associative triples are almost as easy for a person to read and manipulate as English text -- "the X of Y is Z", although stylized, is understandable. ARL resembles a shorthand form we have seen several domain experts use to sketch out sets of rules. The parameter names used in ARL are simply the labels the expert uses in defining the parameters of the domain. The conciseness of ARL makes it much easier to input than English, which is an important consideration when entering a large body of rules. Its conciseness is also a benefit when EMYCIN prints large numbers of rules for the expert to examine. Because knowledge acquisition is a critical problem for building expert systems [cf. Buchanan, 1981], we look on ARL as a pragmatic solution to a large problem we are addressing more fully in a knowledge acquisition system called ROGET.

The debugging package brings together pieces that have been part of the system for some time. These include: 1) the EMYCIN explanation facility; 2) a program that automatically explains how the system arrived at the results of a consultation; 3) a program that reviews each result of a consultation, allowing the user to judge whether the result is correct, and assisting the user in refining the knowledge base in order to correct any errors noted in the result or in intermediate conclusions; and 4) a program that automatically compares the results of a consultation to stored "correct" results for the same case, and explains any errors in the conclusions.

The rule compiler, described in last year's progress report, has been integrated with the whole system. Production rules, while convenient in their modularity, are not the best representation for speedy execution. The rule compiler transforms a program's production rules into a decision tree, eliminating the redundant computation inherent in a rule interpreter, and compiles the resulting tree into machine code. The program can thereby use an efficient deductive mechanism for running the actual consultation, while the flexible rule format remains available for acquisition, explanation, and debugging.

Finally, the EMYCIN user's manual has been improved. This manual is designed to be used by system builders who are creating a consultation system, not by the eventual users of the consultation system itself.

The second major research effort completed this year was an investigation of the efficacy of mixing production rules and frames in a medical reasoning program. This system, called CENTAUR, was described in Jan Aikins' thesis. The medical problem is identical to the topic for the PUFF system, namely the diagnosis of pulmonary function disorders. The medical knowledge, too, was fixed for some of the experiments. The ways the knowledge was represented and used in CENTAUR, however, were changed in order to determine whether improvements over PUFF could be found.

The representation of prototypical cases in CENTAUR improves the understandability of the program's line of reasoning and gives it more focus than PUFF has. Frames (called "prototypes" by Aikins) encode the knowledge of typical cases. These are linked in a hierarchy of ever more specialized descriptions of the subtypes, and are linked to production rules that associate evidence with hypotheses.

The control of reasoning in CENTAUR is also represented in frames, which allows explicit changes in reasoning strategies and experimentations with alternatives. Because of this explicitness, the program can explain its control strategy to a user, thus making it more understandable than in PUFF.

### EMYCIN Applications

EMYCIN is intended for use by system builders who wish to construct a consultation program around a production rule representation of knowledge (with individual facts represented as associative triples) and a backward chaining control structure. It is not a universal programming language,

but because the form of the final consultation program is fixed in advance, EMYCIN can save the system builder considerable effort. We have demonstrated this fact with several applications. Four were described in last year's report (PUFF, SACON, HEADMED, and CLOT). Three more were undertaken in this year, and a fourth was recently begun in the field of dermatology by Dr. Blois and his associates at UCSF.

#### GRAVIDA

A medical consultant called GRAVIDA, was developed to track an expectant mother through her pregnancy. Constructed by Dr. Val Catanzarite, currently a resident at Santa Clara Valley Medical Center, the system acquires information about current and past medical problems of the mother, any previous pregnancies, and general historical data about the patient. GRAVIDA then keeps track of the patient on a per-visit basis, recommending tests, detecting potentially dangerous medical conditions, and estimating the current age of gestation. The construction of this consultant required the extension of the rule language (by adding several new predicate functions) to look for simple trends and events over a series of previous visits.

#### DART

The other consultants are applied to non-medical domains. In conjunction with the IBM Corporation we have developed a consultant, called DART, that identifies probable causes of failures in teleprocessing subsystems of IBM 370-class computer systems. The system accepts stylized descriptions of the observed failure (e.g., lost data, machine went into a loop, terminal doesn't respond, etc.) and then directs the acquisition of data which are collected from traces available to field service personnel. Finally, DART uses these data to indict specific components, both hardware and software, which might be broken.

#### LITHO

The other major consultant now under development seeks to identify rock formations found at various depths of an oil-well bore hole. The consultant, called LITHO, examines geological and physical data of individual zones of interest to identify various aspects of the geological formations. This consultant is being constructed in conjunction with the Schlumberger Corporation and is similar to the GEO consultant developed with the AGE system.

With the publication of van Melle's EMYCIN thesis, which deals with the design of improved knowledge acquisition facilities for EMYCIN, and with the availability of an EMYCIN manual, each of the three consultants described above was constructed largely by the experts themselves. After initial discussions concerning the design of the system's goals, the identification of the data to be gathered, and the basic flow of the consultant's dialogue, the process of writing and inputting the hundreds of rules and parameters per system has been done primarily by the expert. All of them have remarked on the ease with which the current facilities allow

this interaction to occur. As a result of these experiments, numerous improvements and modifications, both to the EMYCIN system and to the manual, are being incorporated into the package.

### GUIDON

The original version of GUIDON, described in Clancey's 1979 thesis, was developed as an experiment to test the educational potential of MYCIN's rules and the ability to use the rule base outside of the consultation setting. Experiments with medical students indicated that GUIDON's framework for teaching knowledge was reasonably satisfactory, but the teaching points were not always clear in the rules. We then conceived a two step plan: first, to analyze the rule set and change it as necessary for teaching purposes (retaining the consultative capability), and second, to test this revised rule set in a new version of GUIDON. In 1980, the first step of this plan was achieved.

### Analysis of MYCIN's Rules

From the period February 1 through December 1, 1980 we met regularly with a physician consultant for the purpose of revising MYCIN's rules so that teaching points were clear. Protocol analysis (presenting cases MYCIN had solved) was the chief method. We also experimented with sorting of medical findings, direct lectures, and undirected recall ("tell me everything you know about..."). We attended a series of courses taught by the same physician and compared them to another physician's handling of the same course.

#### Key Findings:

- a. Our framework of structural, support, and strategic knowledge for organizing, justifying and controlling the use of heuristic rules served well in knowledge acquisition dialogues. We would always ask ourselves, "What kind of rationale is he giving me? A data/hypothesis rule? Why does he believe a rule? Why did he think to consider that association (the indexing, the approach)?" We put our analysis on this psychological footing from the start, because we learned in GUIDON1 that a tutorial program must incorporate knowledge that people use to access and control their heuristics.
- b. It is not sufficient to revise MYCIN's rules; the decomposition of knowledge into subgoals is itself sometimes imprecise or non-standard.
- c. Knowledge has to be added, namely the expertise for when to use MYCIN: when should one think about meningitis? what might it be confused with? MYCIN was not designed to be the "primary care" physician, but teaching diagnosis, our goal, involves expanding the knowledge base to include initial problem formulation.



- d. The physician's approach was logical and easy for us to emulate. He was consistent from case to case, and moreover did what he told students to do. This is not necessarily typical. Other teachers we observed were not able to articulate their approach as clearly and seemed to be less sure of what students were thinking. There were common strategical concepts, however, that our three experts all used to explain their reasoning.

#### Development of NEOMYCIN

We implemented a prototype consultation system that constitutes a psychological model of diagnostic problem solving. This system is upward compatible with EMYCIN systems, and thus could replace the EMYCIN language and interpreter. Key theoretical features of the design:

- a. Forward-directed reasoning from data to hypotheses and state-categories, emulating expert problem solving:
  - 1) Trigger rules place hypotheses in the differential diagnosis directly as data are received. The differential is maintained so that more specific causes replace general hypotheses.
  - 2) Data are abstracted immediately, e.g., "diplopia" is thought of as an "abnormal neurological finding"
  - 3) Process-oriented questions are immediately asked if they are relevant to the domain, even if not directed to any particular hypothesis, e.g., asking when a symptom began and how it has changed over time.
  - 4) Data suggest causal state-categories, possibly jumping over a chain of causal links to conjecture some generic problem.
  - 5) Data/hypothesis associations are applied in the context of the current differential diagnosis (working memory of hypotheses).
- b. Explicit, separate representation of:
  - 1) a problem-space hierarchy to which data/hypothesis rules are attached ("etiological taxonomy") (previously implicit as the "context clauses" of rules);
  - 2) causal rules that ultimately tie into this hierarchy;
  - 3) world relations that constrain the relevance of data (previously implemented as "screening clauses");
  - 4) disease process knowledge that cuts across the etiological distinctions, useful for initial problem formulation.